Computer Science Department

TECHNICAL REPORT

Diagonal Scalings of the Laplacian as Preconditioners for Other Elliptic Differential Operators

A. Greenbaum

Technical Report 488

January 1990

NEW YORK UNIVERSITY

Courant Institute of Mathematical Sciences
251 Mercer Street, New York, N.Y.

Diagonal Scalings of the Laplacian as Preconditioners for Other Elliptic Differential Operators

A. Greenbaum

Technical Report 488

January 1990

Diagonal Scalings of the Laplacian as Preconditioners for Other Elliptic Differential Operators

A. Greenbaum †

Courant Institute of Mathematical Sciences
251 Mercer St.
New York, NY 10012

ABSTRACT

We consider the use of diagonal scalings of the Laplacian matrix as preconditioners for matrices arising from other second order self-adjoint elliptic differential operators. It is proved that if a diffusion operator with a piecewise constant but discontinuous diffusion coefficient is preconditioned by a diagonal scaling of the Laplacian, then, in the limit as the mesh size goes to zero, the optimal diagonal scaling is just the identity. This is in contrast to the case in which the diffusion coefficient is smoothly varying, in which case numerical evidence suggests that the optimal diagonal scaling is approximately equal to the square root of the diagonal of the matrix.

December 2, 1989

† This work was supported by the Applied Mathematical Sciences Program of the US Department of Energy under contract DE-AC02-76ER03077 and by the Advanced Research Projects Agency of the Dept. of Defense under contract F49620-87-C-0065.

Diagonal Scalings of the Laplacian as Preconditioners for Other Elliptic Differential Operators

A. Greenbaum †

Courant Institute of Mathematical Sciences
251 Mercer St.
New York, NY 10012

1. Introduction.

In [2] experiments were reported using a numerical optimization code to determine the preconditioner of a specified form which, for a given coefficient matrix, minimized the condition number of the preconditioned system. One of the more interesting experiments involved finding the optimal diagonal scaling of the Laplacian to use as a preconditioner for other second order self-adjoint elliptic differential operators. Similar experiments had previously been carried out in [1], and the use of preconditioners of this form has also been discussed in [6].

Let A_h be the matrix arising from a finite element or finite difference approximation for the problem

$$\begin{aligned} &-\nabla \cdot a \nabla u = f \quad in\ \Omega \\ &u = 0 \quad on\ \partial\Omega, \end{aligned} \tag{1.1}$$

where the positive coefficient a varies throughout the domain Ω and is bounded away from zero. Let Δ_h be the Laplacian matrix arising from the same finite element or finite difference approximation for the problem

$$\begin{aligned} &-\Delta u = f \quad in\ \Omega \\ &u = 0 \quad on\ \partial\Omega. \end{aligned} \tag{1.2}$$

Let D be any positive definite diagonal matrix. One might consider using the matrix

$$M = D\Delta_h D \tag{1.3}$$

as a preconditioner for the matrix A_h in an iterative algorithm such as the Chebyshev or conjugate gradient method to solve problem (1.1). At each iteration it is then necessary to solve a linear system with coefficient matrix M, but such linear systems are generally much easier to solve than the original problem with matrix A_h. It is trivial to invert the diagonal matrix D, and, on a uniform rectangular grid, Δ_h can be solved with a fast Poisson solver. On an irregular region Δ_h can be solved by embedding the region in a rectangle and using an integral equation formulation of the problem [4]. The number of iterations required by the Chebyshev or conjugate gradient algorithms can be bounded in terms of the condition number of the preconditioned system, and so one might then ask what is the best diagonal matrix D to use in (1.3) in order to minimize this condition number. That is, find a positive definite diagonal matrix D_h such that

† This work was supported by the Applied Mathematical Sciences Program of the US Department of Energy under contract DE-AC02-76ER03077 and by the Advanced Research Projects Agency of the Dept. of Defense under contract F49620-87-C-0065.

$$\min_{D \in \{\text{positive definite diagonal matrices}\}} \kappa((D\Delta_h D)^{-1}A_h) = \kappa((D_h\Delta_h D_h)^{-1}A_h), \tag{1.4}$$

where $\kappa(M^{-1}A_h)$ is the ratio of the largest to smallest eigenvalue of $M^{-1}A_h$, or, the condition number of the symmetrically preconditioned matrix, $M^{-1/2}A_hM^{-1/2}$. This is equivalent to finding a matrix D_h which minimizes

$$\kappa(\Delta_h^{-1}(D^{-1}A_hD^{-1}))$$

over all positive definite diagonal matrices D, since the eigenvalues of $\Delta_h^{-1}(D^{-1}A_hD^{-1})$ are the same as those of $(D\Delta_hD)^{-1}A_h$. The problem was stated in this second form in [1].

In this paper we prove a somewhat counter-intuitive result about the optimal diagonal scaling D_h when the diffusion coefficient a is piecewise constant but discontinuous. Both the result and the method of proof became apparent from studying numerical results of the optimization code, thus indicating the usefulness of such a code as a tool in the study of preconditioners. The result is that in the limit as the mesh size h goes to zero the optimal diagonal scaling D_h approaches the identity (or a scalar multiple of the identity, since scalar factors do not affect the condition number). This is in contrast to the case of a smoothly varying diffusion coefficient a, in which case numerical evidence suggests that the optimal diagonal scaling D_h is approximately equal to the square root of the diagonal of the matrix A_h.

2. A Piecewise Constant Diffusion Coefficient: Theoretical Results.

The first theorem that we prove is very general in nature, applying to arbitrary matrices and preconditioners with a certain algebraic property. It characterizes a space in which the extreme values of the Rayleigh quotient must be attained. The next two theorems use this result and apply to matrices arising from specific forms of equation (1.1), with preconditioners of the form (1.3).

Theorem 1. Let A and C be two n by n symmetric positive definite (SPD) matrices and assume that certain rows of C are just scalar multiples of the corresponding rows of A; that is, there is a nonempty set S such that for each $i \in S$ there is a scalar c_i such that

$$C_{ij} = c_iA_{ij}, \quad \forall\, j=1,\cdots,n. \tag{2.1}$$

Then the extreme values of the Rayleigh quotient $\dfrac{v^TAv}{v^TCv}$ are obtained for vectors v satisfying either

$$(Av)_i = 0 \quad \forall\, i \in S \tag{2.2}$$

or

$$v_j = 0 \quad \forall\, j \notin S. \tag{2.3}$$

Proof. Let w be an arbitrary vector and let v be a vector which satisfies (2.2) and which matches w in all components outside of S. Such a vector exists since A is SPD and hence every principal submatrix is nonsingular. The vector w can be written in the form

$$w = v + \hat{v},$$

where $\hat{v}$ satisfies (2.3). Hence $v^TA\hat{v} = \hat{v}^TAv = 0$ and we have

$$w^T A w = v^T A v + \hat{v}^T A \hat{v}.$$

Since $(Cv)_i$ is also zero for all $i \in S$, it also follows that

$$w^T C w = v^T C v + \hat{v}^T C \hat{v}.$$

Thus the Rayleigh quotient is given by

$$\frac{w^T A w}{w^T C w} = \frac{v^T A v + \hat{v}^T A \hat{v}}{v^T C v + \hat{v}^T C \hat{v}}$$

and since each term in the numerator and denominator is nonnegative, it satisfies

$$\min\left\{ \frac{v^T A v}{v^T C v}, \frac{\hat{v}^T A \hat{v}}{\hat{v}^T C \hat{v}} \right\} \le \frac{w^T A w}{w^T C w} \le \max\left\{ \frac{v^T A v}{v^T C v}, \frac{\hat{v}^T A \hat{v}}{\hat{v}^T C \hat{v}} \right\},$$

from which the desired result follows.

The spaces defined by (2.2) and (2.3) are both A-orthogonal and C-orthogonal to one another, and together they span all of R^n. The technique of expressing a given vector as a sum of vectors from each space will be used in the proofs throughout the paper.

We first use theorem 1 to prove a result about the one-dimensional problem

$$-\frac{d}{dx}(a\frac{du}{dx}) = f, \quad x \in (0,1) \tag{2.4}$$
$$u(0) = u(1) = 0,$$

where the coefficient $a(x)$ has the form

$$a(x) = \begin{cases} a_1, & \text{if } x < .5 \\ a_2, & \text{if } x > .5 \end{cases}, \quad a_1, a_2 > 0, \quad a_1 \neq a_2. \tag{2.5}$$

Let A_h be the matrix arising from a continuous piecewise linear finite element approximation for this problem on a uniform grid of size h. Assume that the grid contains a node at the point of discontinuity of $a(x)$. The matrix A_h is then given by

$$A_h = \begin{bmatrix} 2a_1 & -a_1 & & & & & & \\ -a_1 & & & & & & & \\ & & -a_1 & & & & & \\ & -a_1 & 2a_1 & -a_1 & & & & \\ & & -a_1 & a_1+a_2 & -a_2 & & & \\ & & & -a_2 & 2a_2 & & & \\ & & & & & & -a_2 & \\ & & & & & -a_2 & 2a_2 & \end{bmatrix}. \tag{2.6}$$

Let Δ_h be the one-dimensional Laplacian matrix arising from a continuous piecewise linear finite element approximation for the problem

$$-\frac{d^2u}{dx^2} = f, \quad x \in (0,1) \tag{2.7}$$
$$u(0) = u(1) = 0,$$

on the same uniform grid. The matrix Δ_h is just *tridi* $(-1,2,-1)$. We prove the following theorem.

Theorem 2. Let A_h and Δ_h be as defined above and let D_h be a positive definite diagonal matrix which satisfies (1.4). Then

(1) D_h has the form

$$D_h = \begin{bmatrix} d_{1,h}I_h & & \\ & \bar{d}_h & \\ & & d_{2,h}I_h \end{bmatrix}$$

where $d_{1,h}$, $d_{2,h}$, and $\bar{d}_h$ are positive scalars and I_h is the identity of order $\frac{n-1}{2}$, where $h=\frac{1}{n+1}$.

(2) In the limit as $h \to 0$, these scalars approach each other; that is,

$$\lim_{h\to 0} d_{1,h} = \lim_{h\to 0} d_{2,h} = \lim_{h\to 0} \bar{d}_h \equiv d.$$

If $\hat{D}_h$ is any matrix of the form

$$\hat{D}_h = \begin{bmatrix} \hat{d}_{1,h}I_h & & \\ & \hat{d}_h & \\ & & \hat{d}_{2,h}I_h \end{bmatrix} \tag{2.8}$$

and the positive scalars $\hat{d}_{1,h}$, $\hat{d}_{2,h}$, and $\hat{d}_h$ approach *different* limits as $h \to 0$ (more generally, if there exist positive constants ε and δ such that for all h less than δ either $|d_{1,h} - \hat{d}_h| > \varepsilon$ or $|\hat{d}_{2,h} - \hat{d}_h| > \varepsilon$), then

$$\kappa((\hat{D}_h\Delta_h\hat{D}_h)^{-1}A_h) \geq O(h^{-2}) \quad as \quad h \to 0.$$

We prove this theorem through a series of lemmas. For simplicity we drop the subscript h when it is clear which variables depend on h. The point of discontinuity of a, $x=.5$, is grid point number $\frac{n+1}{2}$, which we denote by m.

Lemma 2.1. Let D be any matrix of the form

$$D = \begin{bmatrix} d_1 I & & \\ & \bar{d} & \\ & & d_2 I \end{bmatrix} \tag{2.9}$$

where d_1, d_2, and $\bar{d}$ are positive scalars and I is the identity matrix of order $m-1$. Define M to be the matrix $D\Delta D$, where Δ is the Laplacian matrix. The vectors v for which the Rayleigh quotient $\frac{v^T A v}{v^T M v}$ attains its extreme values satisfy

$$(Av)_i = 0, \quad i=1, \cdots, m-2, m+2, \ldots, n, \tag{2.10}$$

or, equivalently,

$$\begin{aligned} v_{m-1-j} &= \frac{n-1-2j}{n-1}\, v_{m-1}, \quad j=1, \cdots, m-2 \\ v_{m+1+j} &= \frac{n-1-2j}{n-1}\, v_{m+1}, \quad j=1, \cdots, m-2. \end{aligned} \tag{2.11}$$

Proof. Result (2.10) follows from theorem 1 since all rows of M, except rows $m-1$, m, and $m+1$, are just scalar multiples of the corresponding rows of A and since either of the extreme values, a_1/d_1^2 or a_2/d_2^2, which can be taken on by the Rayleigh quotient for vectors which are zero outside $\{1,\ldots,m-2,m+2,\ldots,n\}$ can also be taken on for vectors satisfying (2.10).

By definition of the matrix A, the equations (2.10) are equivalent to

$$\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & & & \\ & & & & -1 \\ & & & -1 & 2 \end{bmatrix} \begin{bmatrix} v_{m-2} \\ v_{m-3} \\ \cdot \\ \cdot \\ v_1 \end{bmatrix} = \begin{bmatrix} v_{m-1} \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & & & \\ & & & & -1 \\ & & & -1 & 2 \end{bmatrix} \begin{bmatrix} v_{m+2} \\ v_{m+3} \\ \cdot \\ \cdot \\ v_n \end{bmatrix} = \begin{bmatrix} v_{m+1} \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}.$$

It is easy to check that vectors of the form (2.11) are the solutions to these equations, and so, the equivalence of (2.10) and (2.11).

Vectors satisfying (2.10) or (2.11) are called *discrete harmonic*.

Lemma 2.2. (Assertion (2) of the theorem). Let D be any positive definite matrix of the form (2.9). If d_1, d_2, and $\bar{d}$ approach different limits as $h \to 0$, then

$$\kappa(M^{-1}A) \geq O(h^{-2}) \quad \text{as } h \to 0,$$

where $M = D\Delta D$.

Proof: For any vector v satisfying (2.10) and (2.11), we can write

$$\begin{aligned} v^T A v &= v_{m-1}(Av)_{m-1} + v_m(Av)_m + v_{m+1}(Av)_{m+1} \\ &= v_{m-1}a_1(2v_{m-1} - \frac{n-3}{n-1}v_{m-1} - v_m) + v_m((a_1+a_2)v_m - a_1 v_{m-1} - a_2 v_{m+1}) + \\ &\quad v_{m+1}a_2(2v_{m+1} - v_m - \frac{n-3}{n-1}v_{m+1}). \end{aligned}$$

After simplification this becomes

$$v^T A v = a_1 [\, (v_m - v_{m-1})^2 + \frac{2}{n-1}v_{m-1}^2 \,] + a_2 [\, (v_m - v_{m+1})^2 + \frac{2}{n-1}v_{m+1}^2 \,]. \tag{2.12}$$

Similarly, $v^T M v$ can be written as

$$v^T M v = (\bar{d}v_m - d_1 v_{m-1})^2 + \frac{2}{n-1}d_1^2 v_{m-1}^2 + (\bar{d}v_m - d_2 v_{m+1})^2 + \frac{2}{n-1}d_2^2 v_{m+1}^2. \tag{2.13}$$

Taking $v_{m-1} = v_m = v_{m+1} = 1$ and dividing (2.13) by (2.12) gives

$$\begin{aligned} \frac{v^T M v}{v^T A v} &= \frac{(\bar{d}-d_1)^2 + \frac{2}{n-1}d_1^2 + (\bar{d}-d_2)^2 + \frac{2}{n-1}d_2^2}{\frac{2}{n-1}(a_1+a_2)} \\ &\geq \frac{n-1}{2} \cdot \frac{\max\{(d_1-\bar{d})^2, (d_2-\bar{d})^2\}}{a_1+a_2} \geq O(h^{-1}). \end{aligned}$$

Taking $v_{m-1} = \dfrac{\bar{d}}{d_1}$, $v_{m+1} = \dfrac{\bar{d}}{d_2}$, $v_m = 1$, and dividing (2.12) by (2.13) gives

$$\begin{aligned} \frac{v^T A v}{v^T M v} &= \frac{a_1[(1-\bar{d}/d_1)^2 + \frac{2}{n-1}(\bar{d}/d_1)^2] + a_2[(1-\bar{d}/d_2)^2 + \frac{2}{n-1}(\bar{d}/d_2)^2]}{\frac{2}{n-1}(2\bar{d}^2)} \\ &\geq \frac{n-1}{2} \cdot \frac{\max\{a_1(1/\bar{d}-1/d_1)^2, a_2(1/\bar{d}-1/d_2)^2\}}{2} \geq O(h^{-1}). \end{aligned}$$

Hence, by definition of κ, we have

$$\begin{aligned} \kappa(M^{-1}A) &= \left[\max_{v\neq 0}\frac{v^T A v}{v^T M v}\right] / \left[\min_{v\neq 0}\frac{v^T A v}{v^T M v}\right] = \left[\max_{v\neq 0}\frac{v^T A v}{v^T M v}\right] \cdot \left[\max_{v\neq 0}\frac{v^T M v}{v^T A v}\right] \\ &\geq \left[\frac{n-1}{2}\right]^2 \cdot \frac{\max\{a_1(1/\bar{d}-1/d_1)^2, a_2(1/\bar{d}-1/d_2)^2\}}{2} \cdot \frac{\max\{(d_1-\bar{d})^2, (d_2-\bar{d})^2\}}{a_1+a_2} \geq O(h^{-2}). \end{aligned}$$

Lemma 2.3. (Assertion (1) of the theorem.) If D is a matrix of the form (2.9) which satisfies

$$\min_{\hat{D} \text{ of the form (2.9)}} \kappa((\hat{D}\Delta\hat{D})^{-1}A) = \kappa((D\Delta D)^{-1}A)$$

then D also satisfies

$$\min_{\tilde{D}\in\{\text{positive definite diagonal matrices}\}} \kappa((\tilde{D}\Delta\tilde{D})^{-1}A) = \kappa((D\Delta D)^{-1}A).$$

Moreover, any positive definite diagonal matrix which satisfies this equation is of the form (2.9).

Proof: Let $\tilde{D} = diag(\tilde{d}_i)$, $i=1, \cdots, n$ be any positive definite diagonal matrix and let $\hat{D}$ be the matrix of the form (2.9) whose $(m-1)^{st}$, m^{th}, and $(m+1)^{st}$ diagonal elements are equal to those of $\tilde{D}$. Define $\hat{M} = \hat{D}\Delta\hat{D}$ and $\tilde{M} = \tilde{D}\Delta\tilde{D}$. Let v be a vector satisfying (2.10). Then $v^T\tilde{M}v$ satisfies

$$v^T\tilde{M}v = v^T\tilde{D}\hat{D}^{-1}\hat{M}\hat{D}^{-1}\tilde{D}v = w^T\hat{M}w,$$

where $w = \hat{D}^{-1}\tilde{D}v$ matches v in components $m-1$, m, and $m+1$. As in Theorem 1, then, w can be written in the form $w = v + \hat{v}$, where $\hat{v}_{m-1}=\hat{v}_m=\hat{v}_{m+1}=0$, and hence $\hat{v}^T\hat{M}v = v^T\hat{M}\hat{v} = 0$. It follows that

$$v^T\tilde{M}v = w^T\hat{M}w = v^T\hat{M}v + \hat{v}^T\hat{M}\hat{v} \geq v^T\hat{M}v.$$

Since, by theorem 1, the largest value of the Rayleigh quotient $\frac{v^T\hat{M}v}{v^TAv}$ is obtained for a vector v satisfying (2.10), it follows that

$$\max_{v\neq 0}\frac{v^T\tilde{M}v}{v^TAv} \geq \max_{v\neq 0}\frac{v^T\hat{M}v}{v^TAv}. \tag{2.14}$$

Now let a vector w be given by

$$w = \tilde{D}^{-1}\hat{D}v,$$

where v again satisfies (2.10) Then $w^T\tilde{M}w$ is equal to $v^T\hat{M}v$, and, since the $(m-1)^{st}$, m^{th}, and $(m+1)^{st}$ elements of w match those of v, we can again write w in the form $w = v + \hat{v}$, where $\hat{v}_{m-1}=\hat{v}_m=\hat{v}_{m+1}=0$. Hence $\hat{v}^TAv = v^TA\hat{v} = 0$ and we have

$$w^TAw = v^TAv + \hat{v}^TA\hat{v} \geq v^TAv.$$

Since, by theorem 1, the Rayleigh quotient $\frac{v^TAv}{v^T\hat{M}v}$ obtains its largest value for some v satisfying (2.10), it follows that

$$\max_{w\neq 0}\frac{w^TAw}{w^T\tilde{M}w} \geq \max_{v\neq 0}\frac{v^TAv}{v^T\hat{M}v}. \tag{2.15}$$

From (2.14) and (2.15) and the definition of κ, then, the desired result follows:

$$\kappa(M^{-1}A) \le \kappa(\hat{M}^{-1}A) = \left[\max_{v\neq 0} \frac{v^T A v}{v^T \hat{M} v}\right] \cdot \left[\max_{v\neq 0} \frac{v^T \hat{M} v}{v^T A v}\right] \le \kappa(\tilde{M}^{-1}A).$$

Since the inequalities in (2.14) and (2.15) are strict unless $\hat{v}$ is zero, i.e., unless $\tilde{D}$ is, itself, of the form (2.9), the second part of the lemma is also proved.

When the Laplacian Δ_h is used as a preconditioner for A_h, the condition number of the preconditioned system is bounded, independent of h:

$$\kappa(\Delta_h^{-1}A_h) \le \max\{\frac{a_1}{a_2}, \frac{a_2}{a_1}\}.$$

Theorem 2 shows that, for small h, the best diagonal scaling is close to the identity (or a scalar multiple of the identity) and so this bound cannot be improved much. The wrong diagonal scaling can greatly increase this condition number. For larger values of h, however, an appropriate diagonal scaling can significantly reduce the condition number of the preconditioned system, as will be demonstrated in the following section.

A similar result holds for the two-dimensional problem

$$-\left(\frac{\partial}{\partial x}(a\frac{\partial u}{\partial x}) + \frac{\partial}{\partial y}(a\frac{\partial u}{\partial y})\right) = f, \quad (x,y) \in (0,1)x(0,1) \tag{2.16}$$
$$u(x,0) = u(x,1) = u(0,y) = u(1,y) = 0$$

where the coefficient $a(x,y)$ has the form

$$a(x,y) = \begin{cases} a_1, & \text{if } y<.5 \\ a_2, & \text{if } y>.5 \end{cases}, \quad a_1, a_2 > 0, \quad a_1 \neq a_2. \tag{2.17}$$

Again let A_h be the matrix arising from a continuous piecewise linear finite element approximation for this problem on a uniform triangular grid of size h, having a mesh line at the discontinuity, $y=.5$. If the natural ordering of nodes is used then A_h has the form

$$A_h = \begin{bmatrix} a_1T & -a_1I & & & & & & \\ -a_1I & & & & & & & \\ & & -a_1I & & & & & \\ & -a_1I & a_1T & -a_1I & & & & \\ & & -a_1I & \frac{a_1+a_2}{2}T & -a_2I & & & \\ & & & -a_2I & a_2T & & & \\ & & & & & & -a_2I & \\ & & & & & -a_2I & a_2T & \end{bmatrix} \tag{2.18}$$

where $T=tridi(-1,4,-1)$ and I is the identity of order n for a grid of n by n interior nodes. Let Δ_h be the two-dimensional Laplacian matrix arising from a continuous piecewise linear finite element approximation for the problem

$$-\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)=f, \quad (x,y)\in(0,1)x(0,1) \tag{2.19}$$
$$u(x,0)=u(x,1)=u(0,y)=u(1,y)=0$$

on the same uniform grid. The matrix Δ_h is *block tridi*$(-I,T,-I)$. The following theorem is proved very similarly to the 1-D case.

Theorem 3. Let A_h and Δ_h be as defined above and let D_h be a positive definite diagonal matrix which satisfies (1.4). Then

(1) D_h has the form

$$D_h=\begin{pmatrix} d_{1,h}I_h & & \\ & \bar{d}_h I_{5,h} & \\ & & d_{2,h}I_h \end{pmatrix}$$

where $d_{1,h}$, $d_{2,h}$, and $\bar{d}_h$ are positive scalars, I_h is the identity of order $\frac{n(n-1)}{2}$, and $I_{5,h}$ is the identity of order n, where $h=\frac{1}{n+1}$.

(2) In the limit as $h\to 0$, these scalars approach each other; that is,

$$\lim_{h\to 0} d_{1,h} = \lim_{h\to 0} d_{2,h} = \lim_{h\to 0} \bar{d}_h \equiv d.$$

If $\hat{D}_h$ is any matrix of the form

$$\hat{D}_h=\begin{pmatrix} \hat{d}_{1,h}I_h & & \\ & \hat{d}_h I_{5,h} & \\ & & \hat{d}_{2,h}I_h \end{pmatrix} \tag{2.20}$$

and the positive scalars $\hat{d}_{1,h}$, $\hat{d}_{2,h}$, and $\hat{d}_h$ approach *different* limits as $h\to 0$ (more generally, if there exist positive constants ε and δ such that for all h less than δ either $|\hat{d}_{1,h}-\hat{d}_h|>\varepsilon$ or $|\hat{d}_{2,h}-\hat{d}_h|>\varepsilon$), then

$$\kappa((\hat{D}_h\Delta_h\hat{D}_h)^{-1}A_h)\geq O(h^{-2}) \quad as \quad h\to 0.$$

As in the 1-D case, we prove this theorem through a series of lemmas, dropping the subscript when it is clear which variables depend on h. The matrices considered in the 2-D case can be thought of as block matrices, with n blocks, each of order n. The subscript $m=\frac{n+1}{2}$ will denote the middle block, corresponding to the line of discontinuity in a. For any n^2-vector v, v_k will denote the k^{th} block of v.

Lemma 3.1. Let D be any matrix of the form

$$D = \begin{bmatrix} d_1 I & & \\ & \bar{d} I_{.5} & \\ & & d_2 I \end{bmatrix} \tag{2.21}$$

where d_1, d_2, and $\bar{d}$ are positive scalars, I is the identity matrix of order $\frac{n(n-1)}{2}$, and $I_{.5}$ is the identity matrix of order n. Define M to be the matrix $D\Delta D$, where Δ is the Laplacian matrix. The vectors v for which the Rayleigh quotient $\frac{v^T A v}{v^T M v}$ attains its extreme values satisfy

$$(Av)_i = 0, \quad i=1, \cdots, m-2, m+2, \ldots, n, \tag{2.22}$$

Proof: As in the 1-D case, the result is an immediate consequence of theorem 1.

Lemma 3.2. Let v be a vector satisfying (2.22) and let v_{m-1} and v_{m+1} be eigenvectors of $T = tridi(-1,4,-1)$. Then each block v_i can be written in the form

$$\begin{aligned} v_i &= \gamma_i v_{m-1}, \quad i=1,\ldots,m-2 \\ v_i &= \gamma_i v_{m+1}, \quad i=m+2,\ldots,n \end{aligned} \tag{2.23}$$

for some scalars γ_i. If $v_{m\pm1}$ is the eigenvector associated with the smallest eigenvalue of T, then $\gamma_{m\pm2} = 1 - O(h)$.

Proof: Equations (2.22) are equivalent to

$$\begin{bmatrix} T & -I & & \\ -I & & & \\ & & & -I \\ & & -I & T \end{bmatrix} \begin{bmatrix} v_{m-2} \\ v_{m-3} \\ \vdots \\ v_1 \end{bmatrix} = \begin{bmatrix} v_{m-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} T & -I & & \\ -I & T & & \\ & & & -I \\ & & -I & T \end{bmatrix} \begin{bmatrix} v_{m+2} \\ v_{m+3} \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} v_{m+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.24}$$

We will consider only the first set of equations since the second is handled in exactly the same way. Assume there is a solution with $v_i = \gamma_i v_{m-1}$, $i=1,\ldots,m-2$, for some scalars γ_i. Let v_{m-1} correspond to an eigenvalue μ of T. Then equations (2.24) become

$$\begin{bmatrix} \mu & -1 & & \\ -1 & & & \\ & & & -1 \\ & & -1 & \mu \end{bmatrix} \begin{bmatrix} \gamma_{m-2} \\ \vdots \\ \gamma_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.25}$$

This has a unique solution for $\gamma_1, \ldots, \gamma_{m-2}$ if $\mu \geq 2$, and since all eigenvalues of T are greater than 2 this condition holds and the solution of (2.24) is, indeed, of the form (2.23).

Let T_k denote the tridiagonal matrix $tridi(-1,\mu,-1)$ of order k and let $det(T_k)$ denote its determinant. Solving (2.25) using Cramer's rule gives

$$\gamma_{m-2} = \frac{det(T_{m-3})}{det(T_{m-2})},$$

where $det(T_k)$ satisfies

$$det(T_0) = 1, \quad det(T_1) = \mu,$$
$$det(T_k) = \mu\, det(T_{k-1}) - det(T_{k-2}),$$

and $r_k \equiv \dfrac{det(T_k)}{det(T_{k-1})}$ satisfies

$$r_1 = \mu,$$
$$r_k = \mu - \frac{1}{r_{k-1}}, \quad k=2,\ldots,m-2.$$

If carried out indefinitely, this recurrence converges to a solution of the equation

$$r = \mu - \frac{1}{r},$$

namely,

$$r = \frac{\mu + \sqrt{\mu^2-4}}{2},$$

and it is easy to check that after $m-2 = O(h^{-1})$ steps, the ratio r_{m-2} is greater than this limit by $O(h)$. If μ is the smallest eigenvalue of T, then $\mu = 2+O(h^2)$, and so $r_{m-2} = 1+O(h)$ and $\gamma_{m-2} = 1/r_{m-2}$ is $1-O(h)$.

Lemma 3.3. (Assertion (2) of the theorem). Let D be any positive definite matrix of the form (2.21). If d_1, d_2, and $\bar{d}$ approach different limits as $h \to 0$, then

$$\kappa(M^{-1}A) \geq O(h^{-2}) \quad as\ h \to 0,$$

where $M = D\Delta D$.

Proof: Let v be a vector satisfying (2.22), with $v_{m-1} = v_m = v_{m+1}$ being the eigenvector of T, of unit norm, corresponding to the smallest eigenvalue, $\mu = 2+O(h^2)$. Then $v^T A v$ is given by

$$v^T A v = a_1(\mu - 1 - \gamma_{m-2}) + \frac{a_1+a_2}{2}\mu - a_1 - a_2 + a_2(\mu - 1 - \gamma_{m+2}) = O(h)$$

while $v^T M v$ satisfies

$$v^T M v = d_1^2(\mu-\gamma_{m-2})-2d_1\bar{d} + \bar{d}^2\mu + d_2^2(\mu-\gamma_{m+2})-2\bar{d}d_2 > (d_1-\bar{d})^2 + (\bar{d}-d_2)^2 .$$

Hence the ratio satisfies

$$\frac{v^T M v}{v^T A v} \geq O(h^{-1}).$$

If, instead of having unit length, the blocks v_{m-1} and v_{m+1} are taken to have lengths $\frac{\bar{d}}{d_1}$ and $\frac{\bar{d}}{d_2}$, respectively, then we find

$$v^T A v > a_1(\frac{\bar{d}}{d_1}-1)^2 + a_2(\frac{\bar{d}}{d_3}-1)^2$$
$$v^T M v = \bar{d}^2[(\mu-\gamma_{m-2}-1) + (\mu-2) + (\mu-\gamma_{m+2}-1)] = O(h).$$

In this case, then, we have

$$\frac{v^T A v}{v^T M v} \geq O(h^{-1}),$$

and so the condition number satisfies

$$\kappa(M^{-1}A) = \left[\max_{v\neq 0}\frac{v^T A v}{v^T M v}\right] \cdot \left[\max_{v\neq 0}\frac{v^T M v}{v^T A v}\right] \geq O(h^{-2}) .$$

Lemma 3.4. Let $\tilde{D}$ be any positive definite diagonal matrix whose $(m-1)^{st}$, m^{th}, and $(m+1)^{st}$ diagonal blocks are just scalar multiples of the identity. Let $\hat{D}$ be the matrix of the form (2.21) which matches $\tilde{D}$ in blocks $m-1$, m, and $m+1$. Then

$$\kappa((\hat{D}\Delta\hat{D})^{-1}A) \leq \kappa((\tilde{D}\Delta\tilde{D})^{-1}A).$$

Proof: The proof is analogous to that of lemma 2.3 in the 1-D case.

Lemma 3.4 shows that the matrix D of the form (2.21) which minimizes $\kappa((\hat{D}\Delta\hat{D})^{-1}A)$ over all matrices $\hat{D}$ of the form (2.21) also minimizes this quantity over all diagonal matrices whose $(m-1)^{st}$, m^{th}, and $(m+1)^{st}$ diagonal blocks are scalar multiples of the identity. To show that it minimizes this quantity over all diagonal matrices, with possibly nonconstant elements in these blocks, requires some additional work. To this end, we prove the following lemma.

Lemma 3.5. Let D be any positive definite matrix of the form (2.21) and let $M=D\Delta D$. The vectors v for which the Rayleigh quotient $\frac{v^T A v}{v^T M v}$ attains its extreme values have blocks of the form

$$v_i = \alpha_i s, \quad i=1,\ldots,n, \tag{2.26}$$

where $\alpha_1, \cdots, \alpha_n$ are scalars and s is the eigenvector of T corresponding to the smallest eigenvalue.

Proof: Let S be the matrix whose columns are the eigenvectors of T, and let Θ be the diagonal matrix of eigenvalues, so that we have $TS = S\Theta$, $S^TS = SS^T = I$. Define U to be the block diagonal matrix whose diagonal blocks are all equal to S. Then U^TAU is of the form (2.18) with T replaced by Θ and U^TMU is $D^*(block\ tridi(-I,\Theta,-I))^*D$. Let P be a permutation matrix such that the j^{th} column of the i^{th} block of a matrix B is the i^{th} column of the j^{th} block of the matrix BP. Then multiplying by P^T on the left and by P on the right, the above matrices become block diagonal with tridiagonal blocks given by

$$\begin{pmatrix}
a_1\theta_i & -a_1 & & & & & & \\
-a_1 & & -a_1 & & & & & \\
 & -a_1 & a_1\theta_i & -a_1 & & & & \\
 & & -a_1 & \frac{a_1+a_2}{2}\theta_i & -a_2 & & & \\
 & & & -a_2 & a_2\theta_i & -a_2 & & \\
 & & & & -a_2 & & -a_2 & \\
 & & & & & -a_2 & a_2\theta_i &
\end{pmatrix}$$

and

$$\begin{pmatrix}
d_1^2\theta_i & -d_1^2 & & & & & & \\
-d_1^2 & & -d_1^2 & & & & & \\
 & -d_1^2 & d_1^2\theta_i & -d_1\bar{d} & & & & \\
 & & -d_1\bar{d} & \bar{d}^2\theta_i & -\bar{d}d_2 & & & \\
 & & & -\bar{d}d_2 & d_2^2\theta_i & -d_2^2 & & \\
 & & & & -d_2^2 & & -d_2^2 & \\
 & & & & & -d_2^2 & d_2^2\theta_i &
\end{pmatrix},$$

respectively. Clearly, the extreme values of the ratio $w^TP^TU^TAUPw/w^TP^TU^TMUPw$ are attained for vectors w with a single nonzero block, corresponding to the smallest eigenvalue θ_i. The vector $v = UPw$, then, is an extreme vector for the Rayleigh quotient v^TAv/v^TMv, and each block is a scalar multiple of the eigenvector s of T corresponding to the smallest eigenvalue.

Lemma 3.6. (Assertion (1) of the theorem.) If D is a matrix of the form (2.21) which satisfies

$$\min_{\hat{D}\ of\ the\ form\ (2.21)} \kappa((\hat{D}\Delta\hat{D})^{-1}A) = \kappa((D\Delta D)^{-1}A) \tag{2.27}$$

then D also satisfies

$$\min_{\tilde{D}\in\{positive\ definite\ diagonal\ matrices\}} \kappa((\tilde{D}\Delta\tilde{D})^{-1}A) = \kappa((D\Delta D)^{-1}A). \tag{2.28}$$

Moreover, any positive definite diagonal matrix which satisfies (2.28) is of the form (2.21).

Proof: Let $\tilde{D} = diag(\tilde{D}_1, \cdots, \tilde{D}_n)$ be any positive definite diagonal matrix. Let $\hat{D}$ be the matrix of the form (2.21) whose $(m-1)^{st}$, m^{th}, and $(m+1)^{st}$ block coefficients are

$$d_1 = s^T \tilde{D}_{m-1} s, \quad \bar{d} = s^T \tilde{D}_m s, \quad d_2 = s^T \tilde{D}_{m+1} s,$$

where s is the eigenvector of T corresponding to the smallest eigenvalue μ. Define $\hat{M} = \hat{D}\Delta\hat{D}$ and $\tilde{M} = \tilde{D}\Delta\tilde{D}$. Let v be a vector satisfying (2.22) and (2.26). Then $v^T\tilde{M}v$ satisfies

$$v^T\tilde{M}v = v^T\tilde{D}\hat{D}^{-1}\hat{M}\hat{D}^{-1}\tilde{D}v = w^T\hat{M}w,$$

where $w=\hat{D}^{-1}\tilde{D}v$. The vector w can be written in the form $v+\hat{v}$, where $\hat{v} = (\hat{D}^{-1}\tilde{D}-I)v$. Because of the choice of d_1, d_2, and $\bar{d}$, we have

$$\begin{aligned}
\hat{v}^T\hat{M}v = {} & \alpha_{m-1}s^T(d_1^{-1}\tilde{D}_{m-1} - I)^T(d_1^2\alpha_{m-1}\mu s - d_1^2\alpha_{m-2}s - d_1\bar{d}\alpha_m s) + \\
& \alpha_m s^T(\bar{d}^{-1}\tilde{D}_m - I)^T(\bar{d}^2\alpha_m\mu s - d_1\bar{d}\alpha_{m-1}s - \bar{d}d_2\alpha_{m+1}s) + \\
& \alpha_{m+1}s^T(d_2^{-1}\tilde{D}_{m+1} - I)^T(d_2^2\alpha_{m+1}\mu s - d_2^2\alpha_{m+2}s - \bar{d}d_2\alpha_m s) \\
= {} & 0.
\end{aligned}$$

It follows that

$$v^T\tilde{M}v = w^T\hat{M}w = v^T\hat{M}v + \hat{v}^T\hat{M}\hat{v} \geq v^T\hat{M}v.$$

Since, by theorem 1 and lemma 3.5, the largest value of the Rayleigh quotient $\frac{v^T\hat{M}v}{v^TAv}$ is obtained for a vector v satisfying (2.22) and (2.26), it follows that

$$\max_{v\neq 0}\frac{v^T\tilde{M}v}{v^TAv} \geq \max_{v\neq 0}\frac{v^T\hat{M}v}{v^TAv}. \tag{2.29}$$

Now let a vector w be given by

$$w = \tilde{D}^{-1}\hat{D}v,$$

where v again satisfies (2.22) and (2.26). Then $w^T\tilde{M}w$ is equal to $v^T\hat{M}v$, and again we can write w in the form $w = v + \hat{v}$, where $\hat{v} = (\tilde{D}^{-1}\hat{D}-I)v$. We now have

$$\begin{aligned}
\hat{v}^TAv = {} & \alpha_{m-1}s^T(d_1\tilde{D}_{m-1}^{-1} - I)^Ta_1(\alpha_{m-1}\mu s - \alpha_{m-2}s - \alpha_m s) + \\
& \alpha_m s^T(\bar{d}\tilde{D}_m^{-1} - I)^T\left(\frac{a_1+a_2}{2}\alpha_m\mu s - a_1\alpha_{m-1}s - a_2\alpha_{m+1}s\right) + \\
& \alpha_{m+1}s^T(d_2\tilde{D}_{m+1}^{-1} - I)^Ta_2(\alpha_{m+1}\mu s - \alpha_m s - \alpha_{m+2}s),
\end{aligned}$$

which can be written in the form

$$\hat{v}^T A v = v_{m-1}^T (CACv)_{m-1} + v_m^T (CACv)_m + v_{m+1}^T (CACv)_{m+1},$$

where C is a diagonal matrix whose diagonal elements are one except in blocks $m-1$, m, and $m+1$, where they are

$$c_1 = (d_1 s^T \tilde{D}_{m-1}^{-1} s - 1)^{1/2}, \quad \bar{c} = (\bar{d} s^T \tilde{D}_m^{-1} s - 1)^{1/2}, \quad c_2 = (d_2 s^T \tilde{D}_{m+1}^{-1} s - 1)^{1/2},$$

respectively. Because of the choice of d_1, d_2, and $\bar{d}$, we know that the quantities under the square roots are nonnegative, since

$$(s^T \tilde{D}_i s)\cdot(s^T \tilde{D}_i^{-1} s) = (\sum_{j=1}^{n} s_j^2 \tilde{d}_{i,j})\cdot(\sum_{j=1}^{n} s_j^2 \tilde{d}_{i,j}^{-1}) = \sum_{j=1}^{n} s_j^4 + \sum_{j=1}^{n} \sum_{\substack{k=1 \\ k\neq j}}^{n} s_j^2 s_k^2 (\frac{\tilde{d}_{i,j}}{\tilde{d}_{i,k}} + \frac{\tilde{d}_{i,k}}{\tilde{d}_{i,j}})$$

$$\geq \sum_{j=1}^{n} s_j^4 + 2\sum_{j=1}^{n} \sum_{\substack{k=1 \\ k\neq j}}^{n} s_j^2 s_k^2 = (\sum_{j=1}^{n} s_j^2)^2 = 1.$$

Hence $\hat{v}^T A v$ is nonnegative and so we have

$$w^T A w \geq v^T A v + \hat{v}^T A \hat{v} \geq v^T A v.$$

Since, by theorem 1 and lemma 3.5, the Rayleigh quotient $\dfrac{v^T A v}{v^T \hat{M} v}$ obtains its largest value for some v satisfying (2.22) and (2.26), it follows that

$$\max_{w\neq 0} \frac{w^T A w}{w^T \tilde{M} w} \geq \max_{v\neq 0} \frac{v^T A v}{v^T \hat{M} v}. \tag{2.30}$$

Combining (2.29) and (2.30), we obtain the desired result:

$$\kappa(M^{-1}A) \leq \kappa(\hat{M}^{-1}A) \leq \kappa(\tilde{M}^{-1}A).$$

Since the inequalities in (2.29) and (2.30) are strict unless $\hat{v}$ is zero, i.e., unless $\tilde{D}$ is, itself, of the form (2.21), the second part of the lemma is also proved.

While our primary interest has been in *diagonal* scalings of the Laplacian, it should be noted that the proofs of lemmas 2.3 and 3.6 make no use of the assumption that $\tilde{D}$ is diagonal outside of positions (blocks) $m-1$, m, and $m+1$. They can therefore be generalized to the following result:

Corollary. For the 1-D problem, the matrix D_h of theorem 2 minimizes $\kappa((E_h^T \Delta_h E_h)^{-1} A_h)$ over all matrices E_h whose three center rows and columns ($m-1$, m, and $m+1$) have nonzeros only on the diagonal. For the 2-D problem, the matrix D_h of theorem 3 minimizes $\kappa((E_h^T \Delta_h E_h)^{-1} A_h)$ over all matrices E_h whose $3n$ center

rows and columns $(n(m-1)$ through $n(m+1))$ have nonzeros only on the diagonal and these nonzeros are positive. More generally, it minimizes this quantity over all matrices whose $3n$ center rows and columns have nonzeros only in the n by n diagonal blocks and for which these diagonal blocks, E_{m-1}, E_m, and E_{m+1}, have the property that

$$(s^T E_i s)\cdot(s^T E_i^{-1} s) \geq 1, \quad i=m-1,m,m+1,$$

where s is the eigenvector of T corresponding to the smallest eigenvalue.

3. Numerical Results.

For a given matrix A_h an optimization code can be used to determine numerically the optimal preconditioner of the form (1.3). A particularly efficient technique for solving this type of optimization problem was developed by Overton [5]. Experiments with this code were reported in [2]. The code uses a variant of Newton's method to determine the matrix M of a specified form (e.g., form (1.3)) for which the spectral radius

$$\rho(I-M^{-1}A_h)$$

is minimal. It was shown in [2] that this same matrix M (or any scalar multiple of M) also minimizes the condition number $\kappa(M^{-1}A_h)$, provided the set over which the minimization is being performed contains all positive scalar multiples of its members, which it does in this case.

In the following experiment, the matrix A_h was taken to be the matrix arising from a continuous piecewise linear finite element approximation on a uniform grid of size h for the one dimensional problem (2.4)-(2.5), where

$$a_1 = 1, \quad a_2 = 100. \tag{3.1}$$

The optimization code was run to determine the optimal preconditioner of the form (1.3). The diagonal matrix D_h determined by the code was always of the form (2.9), as Theorem 2 shows it must be. (In fact, this observation of the numerical results led to the statement and proof of Theorem 2!) The values of the diagonal elements d_1, d_2, and $\bar{d}$ and the condition number κ of the optimally preconditioned system are listed in Table 1 for various grid sizes. The actual matrix D_h returned by the code has been multiplied by a scalar so that $\bar{d}$ is 1.

The largest problem size that the optimization code was able to handle at the time of these tests was about n=225, but it is currently being modified to handle larger matrices. For this value of n, the element d_1 and the condition number κ have reached only about half of their asymptotic limit. It is not at all clear from the numerical results alone that there is an asymptotic limit, since this is approached only for much smaller values of h. This leads one to question the relevance of asymptotic results such as that in Theorem 2, since for typical size problems they may not be approached. It is interesting to note that for all problem sizes the scalar $\bar{d}$ is equal to d_2, the diagonal element corresponding to the subregion with the larger diffusion coefficient.

$1/h$	d_1	d_2	$\bar{d}$	$\kappa(M^{-1}A_h)$
10	.191	1.000	1	7.22
26	.272	1.000	1	14.69
50	.340	1.000	1	22.89
82	.396	1.000	1	31.10
122	.444	1.000	1	39.08
226	.542	1.000	1	53.50

Table 1. Scalars Defining the Optimal Preconditioner and Condition Number of the Preconditioned System for the One Dimensional Problem (3.1).

Table 2 shows numerical results for a similar two dimensional problem:

$$-(\frac{\partial}{\partial x}a\frac{\partial u}{\partial x}+\frac{\partial}{\partial y}a\frac{\partial u}{\partial y})=f \quad in\ (0,1)x(0,1),$$
$$u(x,0)=u(x,1)=u(0,y)=u(1,y)=0,$$

where

$$a(x,y)=\begin{cases}1, & y<.5\\ 100, & y>.5\end{cases} \tag{3.2}$$

The matrix A_h was again derived from a continuous piecewise linear finite element approximation on a uniform grid of size h. The optimization code was used to find the diagonal matrix D_h for which

$$\kappa((D_h\Delta_h D_h)^{-1}A_h)$$

is minimal, where Δ_h is the 5-point Laplacian.

The matrix D_h was again observed to have the form

$$D_h=\begin{pmatrix} d_1 I & & \\ & \bar{d} I_{.5} & \\ & & d_2 I \end{pmatrix}$$

where d_1, d_2, and $\bar{d}$ are scalars, I is the identity corresponding to the subregion $(0,1)x(0,.5)$ or $(0,1)x(.5,1)$, and $I_{.5}$ is the identity on the dividing line, $y=.5$. The values d_1, d_2, and $\bar{d}$ as well as the condition number κ of the optimally preconditioned system are listed in Table 2 for various grid sizes. Again, since the optimization code was not able to handle very large matrices, the asymptotic behavior of the system cannot be deduced from the numerical results alone, but is established by theorem 3.

$1/h$	d_1	d_2	$\bar{d}$	$\kappa(M^{-1}A_h)$
4	.146	1.251	1	2.30
6	.152	1.143	1	3.23
8	.158	1.076	1	4.07
10	.164	1.034	1	4.85
16	.187	1.000	1	6.93

Table 2. Scalars Defining the Optimal Preconditioner and Condition Number of the Preconditioned System for the Two Dimensional Problem (3.2).

In contrast to the above results, Table 3 shows results for the problem (2.1) where $a(x)$ is given by

$$a(x) = .01 + x^2. \tag{3.3}$$

Although the total variation in $a(x)$ over the interval $(0,1)$ is approximately the same as that in (3.1) ($a_{\max}/a_{\min} = 101$), it now varies smoothly. The diagonal matrix D_h returned by the optimization code is now very nearly equal to the square root of the diagonal of A_h. Table 3 shows the largest and smallest ratio between the square of a diagonal element of D_h and the corresponding element of A_h. D_h has been multiplied by a scalar so that its center element is equal to the square root of the corresponding diagonal element of A_h. In this case, then, the optimal diagonal matrix D_h is not of the form (2.9) and it does not appear to approach the identity as $h \to 0$. Rather it appears to approach the square root of the diagonal of A_h.

$1/h$	$\max_{i=1,\cdots,n} D_{ii}^2/A_{ii}$	$\min_{i=1,\cdots,n} D_{ii}^2/A_{ii}$	$\kappa(M^{-1}A_h)$
10	1.01	.98	1.17
26	1.00	.98	1.26
50	1.00	.99	1.28

Table 3. Ratios of Diagonal Elements for the Optimal Preconditioner and Condition Number of the Preconditioned System for the Problem (3.3).

4. Further Discussion.

In the case of a smoothly varying diffusion coefficient a, the differential operator $\nabla \cdot (a\nabla u)$ can be written in the form

$$a\Delta u + \nabla a \cdot \nabla u \,. \tag{4.1}$$

Consider the equation $\Delta u = -f$. Making a change of variable from u to $v = a^{-1/2}u$ and multiplying this equation by $a^{1/2}$ gives

$$a^{1/2}\,\Delta(a^{1/2}v) \;=\; a\Delta v + \nabla a \cdot \nabla v + a^{1/2}(\Delta a^{1/2})v \;=\; -a^{1/2}f\,. \tag{4.2}$$

The matrix $M = D\Delta_h D$, where D is the square root of the diagonal of A_h, represents the differential operator in (4.2), with the same homogeneous Dirichlet boundary conditions as the original problem. Since this is a second order self-adjoint operator, it follows that using M as a preconditioner for A_h gives a condition

number for the preconditioned system that is $O(1)$, independent of the mesh size [3]. Since the leading terms of the differential operator in (4.2) match those in (4.1), it is perhaps not surprising that this is a near-optimal diagonal scaling.

When the coefficient a is discontinuous or continuous but not differentiable, there is no such analogy between the preconditioner and a differential operator whose leading term(s) match those of the original equation. In this case, a discontinuous diagonal scaling of the Laplacian does not represent a second-order self-adjoint elliptic operator and, as theorems 2 and 3 show for a specific problem class, the condition number of the matrix preconditioned in this way may become infinite as h goes to zero.

References:

[1] P. Concus and G. Golub, "Use of Fast Direct Methods for the Efficient Numerical Solution of Non-separable Elliptic Equations," *SIAM J. Numer. Anal. 10, #6,* 1103-1120, Dec., 1973.

[2] A. Greenbaum and G. Rodrigue, "Optimal Preconditioners of a Given Sparsity Pattern," Courant Institute Technical Report #431, Feb., 1989. To appear in *BIT.*

[3] T. Manteuffel and S. Parter, "Preconditioning and Boundary Conditions," LA-UR-88-2626, Los Alamos Technical Report, July, 1988.

[4] A. Mayo, "The Fast Solution of Poisson's and the Biharmonic Equations on Irregular Regions," *SIAM Jour. Num. Anal. 21, #2,* 285-299, Apr., 1984.

[5] M. Overton, "On Minimizing the Maximum Eigenvalue of a Symmetric Matrix," *SIAM Jour. Mat. An. Appl., Vol. 9, #2,* 256-268, Apr., 1988.

[6] O. Widlund, "On the Use of Fast Methods for Separable Finite Difference Equations for the Solution of General Elliptic Problems," in *Sparse Matrices and their Applications,* ed. by D. J. Rose and R. A. Willoughby, Plenum Press, 1972.

NYU COMPSCI TR-488
Greenbaum, Anne
Diagonal scalings of the
Laplacian as
preconditioners for...c.2